# BIG DATA IN THE OIL AND GAS INDUSTRY:
## A PROMISING COURTSHIP

Michelle R. Tankimovich

TC 660H & PGE 679H
Plan II Honors
Hildebrand Department of Petroleum Engineering
The University of Texas at Austin
May 4, 2018

_____

Dr. Paul Bommer
Department of Petroleum Engineering
Supervising Professor

_____

Dr. John Foster
Department of Petroleum Engineering
Second Reader

## ABSTRACT

Author: Michelle R. Tankimovich

Title: Big Data in the Oil and Gas Industry: A Promising Courtship

Supervising Professor: Dr. Paul Bommer

The energy industry remains one of the highest money-producing and investment industries in the world. The United States' own economic stability depends greatly on the stability of oil and gas prices. Various factors affect the amount of money that will continue to be invested in producing oil. A main disadvantage to the oil and gas industry is its lack of technological adaptation. This weakens the industry because the surest measures are not currently being taken to produce oil in optimally efficient, safe, and cost-effective ways. Big data has gained global recognition as an opportunity to gather large volumes of information in real-time and translate data sets into actionable insights. In a low commodity price environment, saving time, reducing costs, and improving safety are crucial outcomes that can be realized using machine learning in oil and gas operations. Big data provides the opportunity to use unsupervised learning. For example, with this approach, engineers can predict oil wells' optimal barrels of production given the completion data in a specific area. However, a caveat to utilizing big data in the oil and gas industry is that there simply is neither enough physical data nor data velocity in the industry to be properly referred to as "big data." Big data, as it develops, will nonetheless significantly change the energy business in the future, as it already has in various other industries.

## ACKNOWLEDGEMENTS

I would like to thank Dr. Bommer for his knowledge and academic guidance, his patience, and his helpful comments throughout my years in the Petroleum Department. I have learned so much from him. For Dr. Foster, there are not enough words to describe my appreciation for his help. He was my instructor in only my last year at UT, but in that short time he encouraged me to take his graduate course and devoted countless hours to help me hone my knowledge and professional potential. I can only imagine his great influence had he been involved since I started in Petroleum. I am truly indebted to him.

I would also like to thank my Petroleum Engineering advisor, Arletta Tompkins, and my Plan II advisors, Melissa Ossian and Katie O'Donnell, who all have provided strong, supportive guiding hands that helped me navigate the management of two different challenging and very rewarding degrees.

Another thank you goes to my future bosses, Rob Kendall and Sam Newman. By watching their example, they motivate me to give a hundred percent effort a hundred percent of the time. While they wear many hats to be incredibly professionally successful, they still make the time to help everyone in their firm be as successful as they can be.

Lastly, this thesis would not have been possible without the love, support, and encouragement I received from my parents. In the coming years, I hope to show them my deepest gratitude for all they have provided me.

## TABLE OF CONTENTS

## LIST OF FIGURES

## INTRODUCTION

The inspiration for this topic came from witnessing how the oil and gas industry lags behind in key areas of technological advancement as compared to other industries. The problem came into clear relief through direct experience at Houston's Bazean Corporation, a firm that implements machine learning in delegating investment opportunities. The lack of technological advancement refers to the gas and oil industry's failure to use the best software to analyze the value of wells in terms of predicting their production rates, their best completion practices, etc. The better software uses big data through machine learning. Granted, oil and gas has clearly done incredibly well for itself, and this thesis has no intention of discrediting that; rather, it tries to point a way toward helping such a large-capital industry become even more successful through improved efficiency.

Thriving societies, as well as those striving to escape productive stagnancy, need sources of low-cost energy. Low cost energy environments exist when the supply of energy is greater than the demand. Adequate supply exists when sufficient quantities are found and can be made available for consumption in an economically viable way. Finding and producing hydrocarbons, however, is physically challenging and economically risky. In the process of overcoming these challenges and risks, the decisions related to oil and natural gas exploration, development, and production generate large amounts of data – data whose volume grows unimaginably quickly on a daily basis. The industry needs new technologies and approaches to integrate and interpret this data to drive faster and more accurate decisions. This will then lead to safely finding new resources, increasing recovery rates, and reducing environmental impacts.

This thesis explores big data and its impact on various industries in which it already has been developed (e.g. banking, retail, manufacturing, government agencies). Such background can help reveal what big data is, how it works, and the extent to which it is adaptable to – if not indispensable to – the oil and gas industry. This thesis, then, offers a generic version of big data application to oil and gas using a self-designed coding example (there are so many different coding languages and packages within the industry that, without this thesis-specific example, this section could continue indefinitely). Some of the limitations, challenges, and benefits of this coding example, will be presented. The thesis concludes speculating about why big data has only recently garnered serious attention and implementation in oil and gas, while so many other industries have used it already for quite some time. The main conclusion is that big data and the oil and gas industry are in a necessary courtship that has the potential to blossom into a long and productive marriage.

## BACKGROUND

### What is Big Data?

Big data is a term that describes extremely large volumes of data that are analyzed computationally to expose trends, patterns, and associations. The concept is continually evolving and being continually re-envisioned. Big data is the driving force behind ongoing waves of digital transformation, such as artificial intelligence and data science (Marr, n.d. -a). The amount of data one has is important, as more data leads to more understanding of the entire mechanism at hand. However, the important takeaway of big data is what organizations do with the data, as analyzed data provides insights that lead to better tactical decisions and strategic business moves (Big data, n.d.).

Big data needs to be evaluated with five main components in mind – volume, velocity, variety, variability, and complexity. Volume of data signifies the quantity of data or the size of the data set. Velocity of data is the speed at which the data is retrieved and streamed. Variety of data refers to receiving structured, numeric data in traditional databases all the way to unstructured documents, videos, audio, etc. Variability of data is the ability to manage inconsistent data flows, which increase in inconsistency with higher velocity levels and more data varieties, structured and unstructured. Complexity of data denotes the multiple sources from which data originates and the difficulties in linking, matching, cleaning, and transforming data across different systems in order to connect and correlate relationships between the multiple sources (Big data, n.d.).

Due mainly to the rise of computers, the Internet, other electronic networks, and thus the overall collected data, big data agglomerates quickly, increasing data volume

and complexity at an obscenely high rate. It is important to note that data, in itself, is not a new invention because even before computers, people were collecting data in records, archives, etc. The important takeaway though is that computers, spreadsheets, and databases, gave a way to store and organize data on a massive scale and in an easily accessible manner (Marr, n.d. -a).

Technology is now well beyond spreadsheets and databases. For example, "today, every two days we create as much data as we did from the beginning of time until 2000" (Marr, n.d. -a). Almost every action taken leaves a digital trail. A digital footprint is left with almost every action that everyone performs, as most actions nowadays require a digital aspect to them. In addition, the amount of machine-generated data is rapidly growing too, in terms of simply "smart" home devices communicating to home servers and all the way to industrial machinery sensors gathering and transmitting data (Marr, n.d. -a).

**How Does Big Data Work?**

In order for big data to work properly, it continually relies on more information about everything and anything. Big data processes have an infinite appetite, but instead of getting fat, big data becomes leaner. With more data, more reliability and accuracy come with new insights, predictions, and projections for the future. "By comparing more data points, relationships begin to emerge that were previously hidden," and these relationships enable businesses to make smarter decisions (Marr, n.d. -a). This is done mainly through building models, based on the collected data, and then running simulations. The data points can be tweaked each time and then monitored to see how those changes impacts the results. This is an automated process. Analytics technology runs millions of these simulations, tweaking all the possible

variables until it finds a pattern, or an insight, that helps solve the problem (Marr, n.d. -a).

The sources for big data generally fall into one of three categories: streaming data, social media data, or publicly available sources. Streaming data consists of data that reaches IT systems from a web of connected devices. Social media data includes the data on social interactions. This is an attractive set of material, particularly for marketing, sales, and support functions. It is often in unstructured or semi structured forms, which creates a challenge to analysis. Publicly available sources include the data available from open data sources (Big data, n.d.).

Advances in storage and analytics allow companies to capture, store, and work with many different types of data sets, not just ones that come in perfectly organized spread sheets or databases. In order to clean, or to make sense of unorganized data, big data projects use cutting-edge analytics, involving artificial intelligence and machine learning (Marr, n.d. -a).

Artificial Intelligence (AI) is "'the field of computer science dedicated to solving cognitive problems commonly associated with human intelligence, such as learning, problem solving, and pattern recognition'" (Marr, 2018). Machine learning is the leading edge of AI. Most simply, it is about teaching computers to learn in the same manner humans do, by interpreting data, classifying it, and learning from its successes and failures. The foundation of machine learning is building algorithms around the binary "yes" and "no" logic of computers that are capable of basically processing data to the level of the human mind. Moving on from this basic foundation, businesses really capitalize on machine learning when they practice "deep learning." Deep learning is derived from "'deep' neural nets" which are many layers of multiple

networks built on top of each other, "passing information down through a tangled web of algorithms to enable a more complex simulation of human learning" (Marr, n.d. -b).

By implementing AI, companies teach computers how to identify what a particular set of data represents – for example, through either image recognition or natural language processing. Computers quickly learn to spot patterns much faster and more reliable than humans (Marr, n.d. -a). Thus, using AI with an incredibly large amount of data, allows different sectors, like businesses, governments, health care providers, to implement big data. Big data then creates tremendous amounts of knowledge to be used towards running operations more efficiently and creating decisions through predictions.

## THE IMPORTANCE OF BIG DATA

The main important aspect of big data is not necessarily the volume of the data, but what actions can be performed on the data and then the gained conclusions. Big data answers problems that find solutions to cost reductions, time reductions, new product development and optimized offerings, and smart decision making (Big data, n.d.).

At an unprecedented rate, data is changing so many aspects of the world and thus affect the way people live in it. The amount of data available only increases and analytics technology will only advance. "For businesses, the ability to leverage Big Data is only becoming increasingly critical" and those that view data as a strategic asset are the ones that will survive and thrive, but those that ignore the big data phenomena will risk their ability to succeed (Marr, n.d. -a).

Big data affects organizations in almost every industry, such as banking, education, government, health care, manufacturing, and retail. Each of these can benefit greatly through the use of big data. In banking, big data brings insights into customer care, innovative ways to manage data, minimizing risk and fraud. Big data in education can lead educators to identify at-risk students and to implement better systems to evaluate teachers. When government agencies apply big data analytics, insights arise into "managing utilities, running agencies, dealing with traffic congestion or preventing crime" (Big data, n.d.). However, governments must address the issues of transparency and privacy to its citizens. With big data, health care providers can discover unknown discernments that will improve patient care. Manufacturers can boost quality and output while minimizing waste. Retailers can find out the best way

to market to customers, the most effective way to handle transactions, and the most strategic way to bring back lapsed business" (Big data, n.d.).

An example of the benefit of big data is the case study of UPS. UPS stores a large amount of data, much of which comes from sensors in its vehicles. When this data was analyzed, a major redesign of UPS drivers' route structures occurred. The initiative was called ORION (On-Road Integration Optimization and Navigation). ORION was arguably the world's largest operations research project and it relied heavily on online map data to reconfigure a driver's pickups and drop-offs in real time. The project led to savings of more than 8.4 million gallons of fuel by cutting 85 million miles off of the daily routes. UPS estimates that saving only one daily mile per driver saves the company $30 million. This is just one of the many substantial solutions from big data (Big data, n.d.).

Amazon builds many of its business on machine-learning systems. Machine learning is so important to Amazon, they stated, "'without ML, Amazon.com couldn't grow its business, improve its customer experience and selection, and optimize its logistic speed and quality'" (Marr, 2018).

Based on Bernard Marr's research, machine and deep learning are the priority for Google AI and its tools to "'create smarter, more useful technology and help as many people as possible'" from translations to healthcare to making smartphones even smarter (2018). Facebook AI is committed to "'advancing the file of machine intelligence and are creating new technologies to give people better ways to communicate'" (Marr, 2018). IBM's three areas of focus include AI Engineering, AI Tech, and AI Science. AI Engineering builds scalable AI models and tools. AI Tech creates the core capabilities of AI, such as natural language processing, speech and image recognition and reasoning. AI Science focuses on expanding the frontiers of AI.

In 2016, several industry leaders including Amazon, Apple, DeepMind, Google, IBM, and Microsoft joined together to create Partnership on AI to Benefit People and Society in order to develop and share the best practices, to advance public understanding, to provide an open platform for discussion, and to identify aspirational effort in AI for socially beneficial purposes (Marr, 2018).

According to Marr, those who work with AI, make it a priority to define the field for the problems it will solve and the benefits the technology can have for society. The main objective is no longer to achieve AI that operates just like a human brain, but to use AI's unique capabilities to enhance the world (2018).

Big Data is revolutionizing the world across almost every industry to various degrees. Efficiency levels of companies' operations are drastically increased. In addition, "companies can now accurately predict what specific segments of customers will want to buy, and when, to an incredibly accurate degree" (Marr, n.d. -a). Even outside of business, big data projects help change the world in a number of ways, such as in the healthcare, weather, and crime sectors, to name just a few.

In the healthcare industry, data-driven medicine involves analyzing vast numbers of medical records and images for patterns that can help spot diseases early and develop new medicines. In the weather sector, big data projects lead to predicting and responding to natural and man-made disasters. Sensor data is analyzed to predict where earthquakes are likely to strike next. In addition, analyzing patterns of human behavior, give clues that help organizations give relief to the survivors. Big data can also help prevent crime as "police forces are increasingly adopting data-driven strategies based on their own intelligence and public data sets in order to deploy resources more efficiently and act as a deterrent where one is needed" (Marr, n.d. -a).

On a much larger scale, people use big data technology to monitor and safeguard the flow of refugees away from war zones around the world.

## BIG DATA WITHIN THE OIL AND GAS INDUSTRY

### Big Data's Potential

For decades now, the oil and gas industry has been gathering data and automating its systems. Combining the institutional knowledge of the energy domain and the data collected from equipment and well fields, analytics, AI, machine learning – and, therefore, big data – helps industries to formulate solutions in a structured setting. In addition, using big data, businesses increase efficiency levels, which then improves productivity levels and reduces unplanned problems (Gruss, 2018).

Big data in the energy domain was only adopted recently, unlike other industries. Within energy analytics, only small-scale data analysis was the initial step. This could only compare "individual or a small number of equipment in a production line or refinery at a time" (Gruss, 2018). Nowadays, the process needs to be scaled across the globe. This is a challenge, yet also an opportunity at the same time. If it is globally scaled, companies could "analyze turbines across all their units at one go and normalize operations based on different factors like weather, operating conditions, or production medium (oil, water, and gas)" (Gruss, 2018).

According to Dr. Alec Gruss and Robert Skiebe, both project directors at the renowned Siemens Power and Gas, the following is an example of implementation of big data in the energy industry:

> Siemens has been doing vibration analysis to detect equipment faults and downtime as a component of their condition monitoring and predictive maintenance offerings for close to two decades now. Thus the company has the pre-requisite to understand what an enterprise needs to do in terms of its data collection and processing to make the energy equipment data available at a large scale, much more efficiently to conduct in-depth analytics. (2018).

Businesses are forced to examine their energy usage, as energy prices are expected to remain elevated. They do this with the intention of cutting costs to stay competitive and / or generate value from on-site energy assets. ERM Business Energy is the second largest commercial and industrial energy retailer in the country and delivers tailored solutions across a range of businesses and sectors, such as state and federal government agencies and some of the country's largest businesses. According to ERM Business Energy, the most cost-effective way to decrease costs is to use data analytics in order to manage energy costs with an end-to-end solution. This solution can fund itself in a short period of time, the upside for a business.

According to ERM's executive general manager Megan Houghton says:

> In developing the best integrated solution for a business, data is king. Any solution should start with a full and transparent energy audit so businesses can understand how they're using energy, identify areas of energy waste and which areas can be targeted for savings. (Using energy, 2017).

This audit provides an entire analysis of how a business is using energy – where the biggest spends are and how buildings, plants, and equipment are consuming energy. This allows evidence-based decisions and then prioritization of energy efficiency actions. Houghton stresses the importance of undertaking energy initiatives in the right order. This ensures the best solution and the best use of capital. Data analytics play an increasingly important role in energy management. Data analytics not only identifies the problem, but aides in designing the best value solution and monitoring the outcome (Using energy, 2017).

According to ERM's digital innovation and advisory lead Peter Tickler, "investing in big data and analytics makes your energy productivity decisions more

accurate, targeted and prioritized" (Big data, 2018). Big data allows for predictive models that determine future outcomes. These highly sophisticated data models come from data scientists overlaying millions of data points, originating from varied data sets. These rich and accurate data models are used to make informed and evidenced-backed investment decisions. Big data is leading to more automation in terms of the energy industry's consumption and management, thus creating informed energy productivity decisions for both now and into the future. Big data is at the heart of innovation for the industry (Big data, 2018).

One of the primary assets of successful, thriving societies is a low-cost energy source. Low cost environments exist when the supply is greater than the demand. Supply exists when sufficient quantities are found, so producing oil and gas is economically viable. The processes and decisions related to oil and natural gas exploration, development, and production generate large amounts of data and the data volume grows daily. Finding and producing hydrocarbons is challenging and economically risky. The industry needs new technologies and approaches to integrate and interpret this data to drive faster and more accurate decisions. This will then lead to safely finding new resources, increasing recovery rates, and reducing environmental impacts (Farris, 2012).

This data creates models and images of Earth's structure and layers 5,000-35,000 feet below the surface. In addition, the models describe activities around the wells themselves, such as machinery performance, oil flow rates, and pressures. With approximately one million wells currently producing oil and /or gas in the United States alone, this dataset is growing daily (Farris, 2012).

Adam Farris is the senior vice president of business development for the company Drillinginfo. Drillinginfo is a leading data and intelligence provider of upstream data for oil and gas decisions. According to Farris, the need for big data analysis and its potential for reward are great. More than 20,000 companies are associated with the oil business. Almost all of them need data analytics and integrated technology throughout the oil and gas lifecycle. The next decade must focus on ways to use all of the data the industry generates in order to automate simple decisions and guide harder ones. This all would ultimately reduce risk and result in finding and producing more oil and gas with less environmental impact (2012).

However, finding and developing oil and gas while also reducing safety risk and environmental impact is a difficult task. The main three big oil and gas industry problems that consume money and produce data are discovery, drilling, and producing oil and gas. The layers of hydrocarbon-bearing rock are deep below the Earth's surface and most hydrocarbons are locked in hard-to-reach places, such as in deep water or areas with difficult geopolitics. Oil is stored in the tiny pores between the grains of rock, and much of the oil-containing rock oil is extremely tight. Further, oil is found in areas that have structurally trapped the oil and gas, and, without a structural trap, oil and gas can migrate throughout the rock, resulting in lower pressures and uneconomic deposits. In addition, each reservoir has its own unique recipe to get the most quantity out of the ground profitably and safely. The path to optimizing production is dependent on the type of rock and structure of the reservoir. These decisions depend heavily on models made by the foundation of available data (Farris, 2012).

According to Farris, "analytical approaches that impact the success rate of finding or reducing the cost to develop and produce oil and gas can make energy more affordable, safer and environmentally conscious" (2012). Data science will help the oil

and gas industry learn more about each subsystem and create more accuracy and confidence in every decision, ultimately reducing risk. Advanced predictive methods with self-learning capabilities to use new and previously unknown or unavailable data will truly be the breakthrough technology that provides better insight into process and asset dynamics. In addition, the oil and gas industry has to manage risks on many fronts, and the assets employed are expensive and capital intensive. Any systems supporting the industry must be highly reliable, responsive and secure. In addition, many of the assets are geographically widely dispersed. Taking all of these considerations in mind, the computational resources needed to support this industry reveal to have greater demands than typical commercial computing platforms. Big data analytics is the answer to these demands (Miklovic, 2017).

While the use of big data is still in its infancy, as far as the oil and gas industry is concerned, there are still some possible near-term big data analytical solutions. One of them is the integration over a wide variety of large data volumes. This allows finding additional hydrocarbons, and identifying the data and the best-known technologies to produce it. Another solution is to make daily operational data relevant, in order to reduce operating costs, improve recovery rates, and decrease environmental concerns. Another near-term solution is increasing the quality of decision management. This takes into account everything that is known and quickly identifies if or how to proceed. As oil and gas companies awake to the potential of analytics, many jobs will be created for data scientists, opening a portal for new applications and ideas to enter the industry (Farris, 2012).

**Ways to Process Big Data**

Effective data-driven science and computation requires an understanding of how data is stored and manipulated. There are countless coding languages, packages,

libraries that can be used on big data. The following sections provide only a generic glimpse into a few ways the oil and gas industry implements big data.

The following includes some basic definitions of different aspects within the coding language Python. At the basic level is a module, which is a file that contains Python functions and global variables. It is a file that defines one or more functions / classes, which are intended to be re-used in different codes of the program. At the next level is a package, which is a directory of Python modules. The next level is a library, which is a collection of various packages. There is no conceptual difference between a package and Python library.

Datasets are heterogeneous, as they can come from a wide range of sources and a wide range of formats, including collections of documents, collections of images, collections of sound clips, collections of numerical measurements, or nearly anything else. Despite these varied forms, data should be thought of as arrays of numbers. Images, sound clips, text, etc. can all be converted into various numerical representations, such as binary digits. The first step in making data analyzable is to transform the data into arrays of numbers. Because of this, efficient storage and manipulation of numerical arrays is absolutely fundamental to the process of doing data science. Python's specialized tools for handling such numerical arrays are the NumPy package and the Pandas package (VanderPlas, 2016c).

Within Python, there are different packages that provide different applications and manipulation techniques for data stored within them. One example is NumPy and its specific "ndarray" function, which provides efficient storage and manipulation of dense typed arrays. Numerical Python (NumPy) provides an efficient interface to store and operate on dense data buffers, including useful storage and data operations as the arrays grow larger in size. NumPy arrays form the core of nearly the entire ecosystem

of data science tools in Python (VanderPlas, 2016c). Another type of library within Python is Pandas. Pandas is a newer package whose data structures allow an efficient implementation of a "DataFrame." DataFrames are essentially multidimensional arrays with attached row and column labels, and are often with heterogeneous types of data and / or missing data. Pandas offers a convenient storage interface for labeled data and implements a number of powerful data operations, familiar to users of both database frameworks and spreadsheet programs (VanderPlas, 2016a).

NumPy's "ndarray" data structure provides essential features for the type of clean, well-organized data typically seen in numerical computing tasks. While it serves this purpose very well, it lacks capability in flexibility within the data (e.g., attaching labels to data, working with missing data, etc.) and with analyzing less structured data. In the real-world data is rarely clean and homogeneous. In particular, many interesting datasets will have some amount of data missing. To make matters even more complicated, different data sources may indicate missing data in different ways. Pandas, and its "Series" and "DataFrame" objects, builds on the NumPy array structure and provides efficient access to tasks that occupy much of a data scientist's time (VanderPlas, 2016a).

At the very basic level, Pandas objects can be thought of as enhanced versions of NumPy structured arrays in which the rows and columns are identified with labels rather than simple integer indices. Pandas provides a host of useful tools, methods, and functionality on top of the basic data structures. Keeping the context of data and combining data from different sources – both potentially error-prone tasks with raw NumPy array – become essentially foolproof tasks with Pandas. The three fundamental Pandas data structures are the "Series," the "DataFrame," and the "Index" (VanderPlas, 2016b).

Machine learning is the primary means by which data science manifests itself in the broader world. It is where these computational and algorithmic skills of data science meet the statistical thinking of data science. The result is a collection of approaches to inference and data exploration that are not about effective theory so much as effective computation. While these machine learning methods can be incredibly powerful, to be effective they must be approached with a firm understanding of the strengths and weaknesses of each method, as well as an understanding of general concepts such as bias and variance, overfitting and underfitting, and more (VanderPlas, 2016d).

Machine Learning is a means of building mathematical models to help understand data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. In this way, the program can be considered to be learning from the data. This type of mathematical, model-based learning is similar to the learning exhibited by the human brain (VanderPlas, 2016e).

At the most fundamental level, machine learning can be categorized into two main types: supervised learning and unsupervised learning. Supervised learning involves modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into classification tasks and regression tasks. In classification, the labels are discrete categories, while in regression, the labels are continuous quantities. Unsupervised learning involves modeling the features of a dataset without reference to any label and is often described as "letting the dataset speak for itself." These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct

groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. In addition, there are semi-supervised learning methods, which fall somewhere between supervised learning and unsupervised learning. Semi-supervised learning methods are often useful only when incomplete labels are available (VanderPlas, 2016e).

There are several Python libraries which provide solid implementations of a range of machine learning algorithms. One of the best known is Scikit-Learn, a package providing efficient versions of a large number of common algorithms (VanderPlas, 2016e).

The following is a supervised learning example of a regression task, in which the labels are continuous quantities. With supervised learning, we are trying to build a model that will predict labels for new data. Consider the data shown in the following figure, which consists of a set of points each with a continuous label:

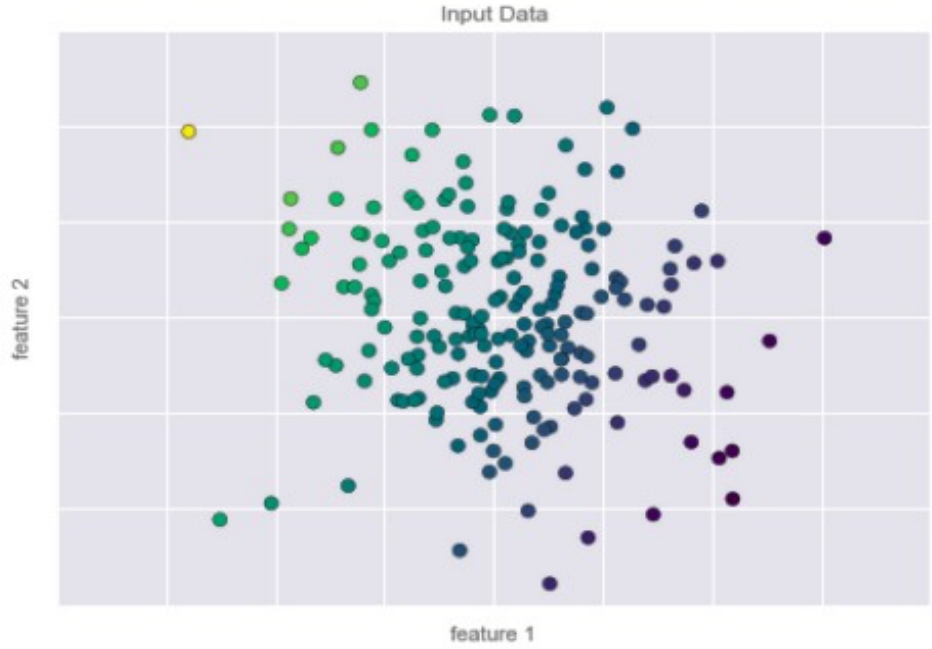


*Figure 1*. Initial 2D Data (VanderPlas, 2016f)

According to Dr. VanderPlas, this is two-dimensional data, that is, there are two features describing each data point. The color of each point represents the continuous label for that point. A simple linear regression model is used to predict the points, assuming that if treating the label as a third spatial dimension, then a plane can be fitted to the data. This is a higher-level generalization of the well-known problem of fitting a line to data with two coordinates (2016e). This setup is shown in the following figure:

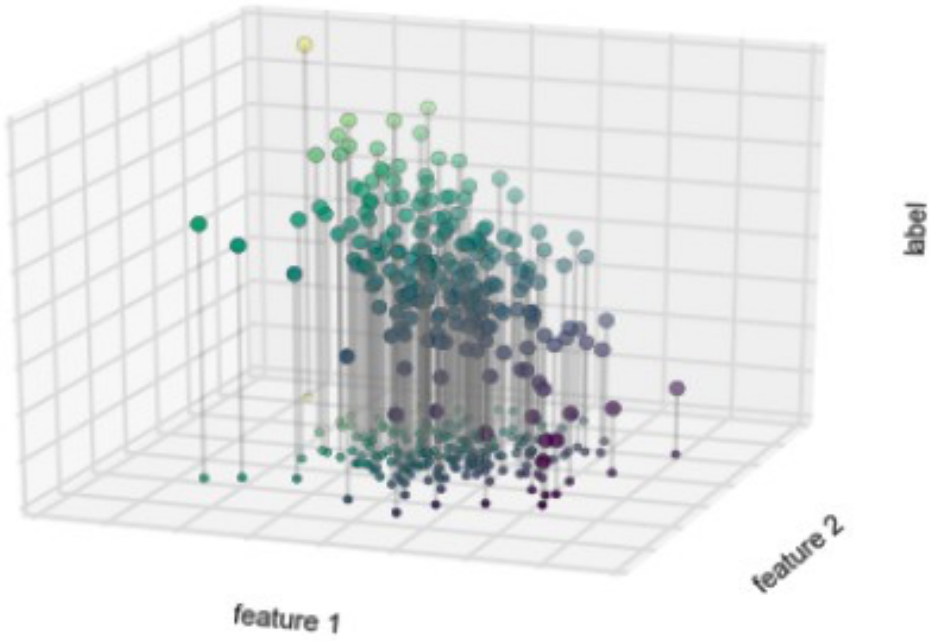


*Figure 2*. Data Seen with Added Third Spatial Dimension (VanderPlas, 2016f)

The feature 1—feature 2 plane here is the same as the two-dimensional plot seen in the first image. However, here the labels are represented by both color and a three-dimensional axis position. From this view, it seems reasonable that fitting a plane through this three-dimensional data would allow to predict the expected label for any set of input parameters (VanderPlas, 2016e). Returning to the two-dimensional projection, when fitting such a plane, this is the result:

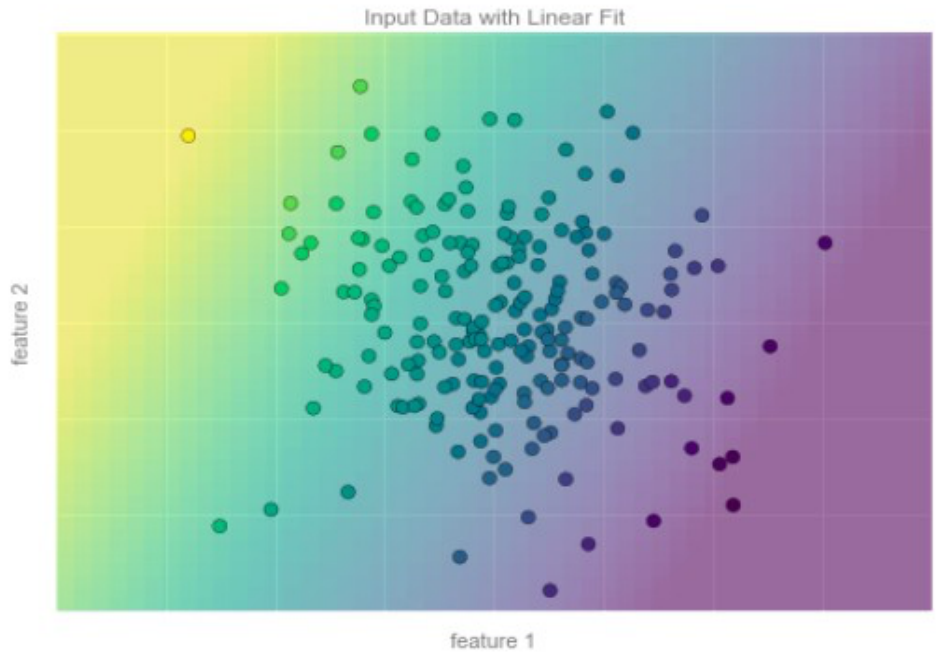


*Figure 3*. Fitting Plane to 2D Projection (VanderPlas, 2016f)

This plane of fit helps to predict labels for new points. An example of this:

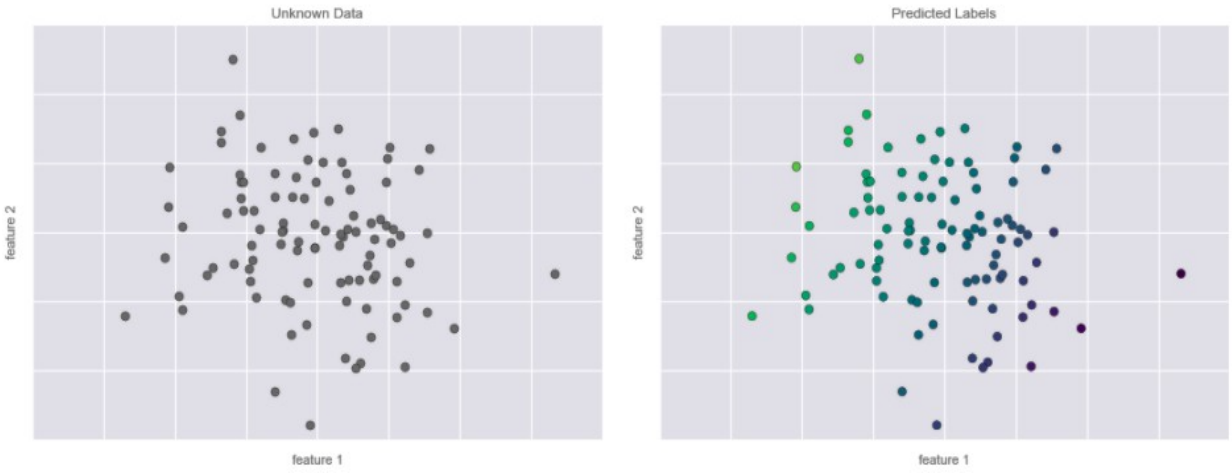


*Figure 4*. Prediction for New Points (VanderPlas, 2016f)

This may seem rather trivial in a low number of dimensions; however, the power of these methods is that they can be straightforwardly applied and evaluated in the case of data with many features (VanderPlas, 2016e).

According to Dr. VanderPlas, unsupervised learning involves models that describe data without reference to any known labels. Within it, one common case is clustering, where data is automatically assigned to some number of discrete groups (2016e). For example, here is some two-dimensional data:

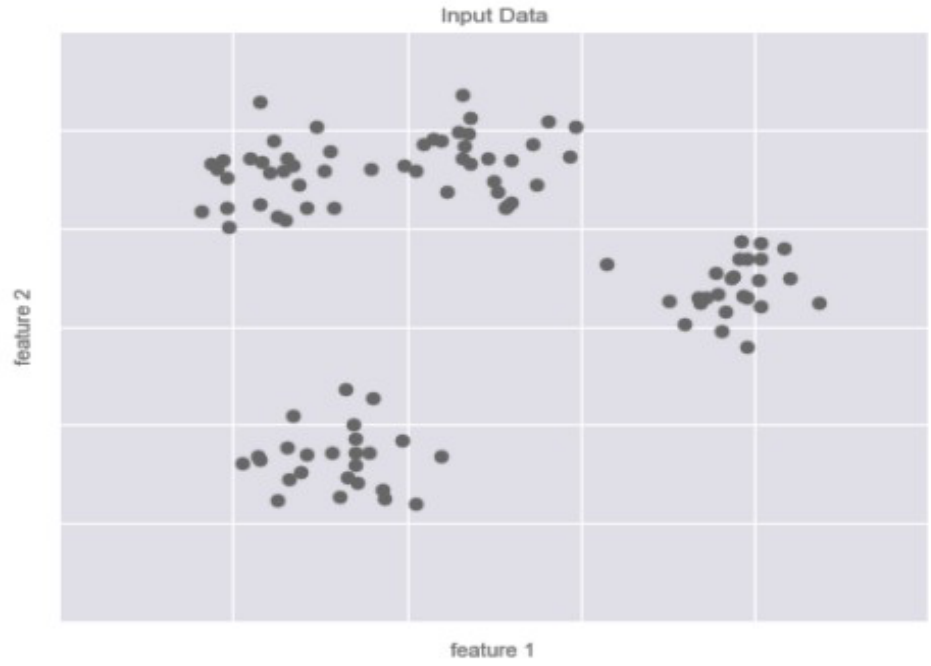


*Figure 5*. Initial 2D Data (VanderPlas, 2016f)

By eye, it is clear that each of these points is part of a distinct group. Given this input, a clustering model will use the intrinsic structure of the data to determine which points are related (VanderPlas, 2016e). Using a very fast and intuitive algorithm (*k*-means), the found clusters are shown:

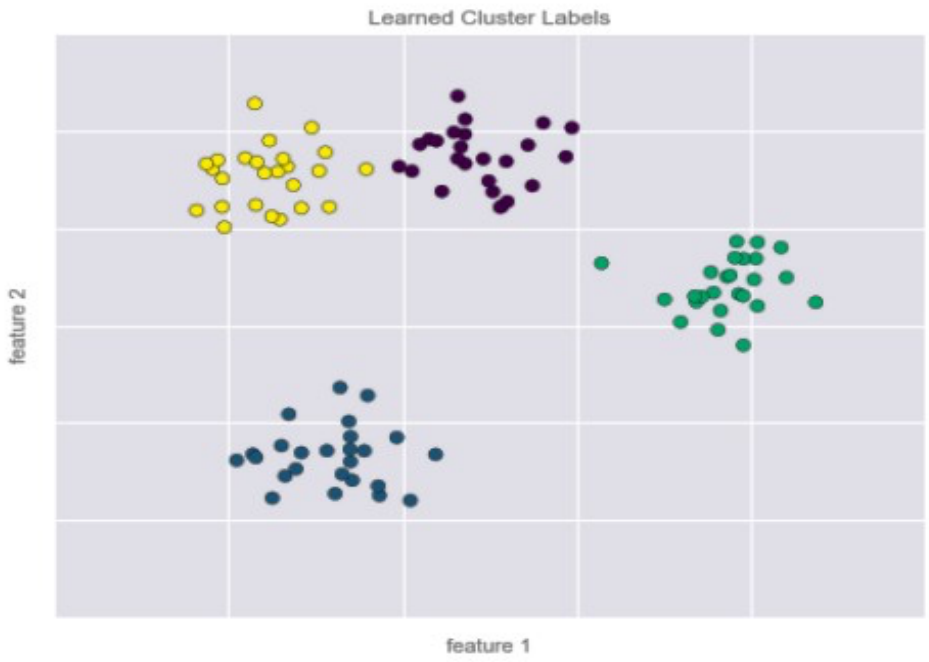


*Figure 6*. Clustered Data (VanderPlas, 2016f)

The *k*-means algorithm fits a model consisting of *k* cluster centers. The optimal centers are assumed to be those that minimize the distance of each point from its assigned center. This might seem like another trivial exercise in two dimensions, but as data becomes larger and more complex, such clustering algorithms can be employed to extract useful information from the dataset (VanderPlas, 2016e).

These were just a couple of broad examples to give an initial background to the practice of machine learning. The next section shows my own example, where I personally implemented supervised learning with my own code.

**A Personal Case Using Supervised Learning**

The following shows a case where supervised learning was implemented. The goal was to show the high level of accuracy for predictions of wells' oil production, based off of machine learning through linear regression.

Bazean provided an excel sheet of fifty wells in New Mexico within the Permian basin. This excel file contained fifty rows of each well and then about sixty columns of information ranging from production of oil in barrels, completion values, fracture concentrations, fracture proppants, number of producing days, permit dates, operator names, etc. The following paragraph includes the steps which were taken to produce a graph that shows the accuracy of predicted values. This is just a generic explanation of some of the code, not all the steps of the code are thoroughly explained. As noted in the introduction, had all of the codes been addressed, this section would become convoluted and take away from the meaning of the produced graph and objective of this code, in my opinion.

The first step was to run this excel file into Python, through Jupyter Notebook. Jupyter Notebook is an open-source web application that allows the creating and sharing of documents that contain live code, equations, visualizations and narrative text. Some examples of use include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more (Jupyter, 2018). The next step was to convert all text labels into binary data, to fill missing data in any cells with "not a number" (NaN), and then input the NaNs with the specific column's average value. This was all done with "DataFrameMapper," it allows for mapping pandas data frame columns into different transformations, to put it in a very simple manner. Not enough information about that data was given to determine a better or more suitable approach other than to replace missing data with average values. This can be seen in Figure 7.

```
import sklearn
import numpy as np
import pandas as pd

#Save my excel file into csv and run into python
df = pd.read_csv("/Users/michelletankimovich/Desktop/new_well_data.csv")

from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import Imputer, LabelBinarizer, PolynomialFeatures, StandardScaler
from sklearn.linear_model import Ridge, LassoCV, LinearRegression
from sklearn.model_selection import train_test_split
from sklearn_pandas import DataFrameMapper

#Assigning an index column to pandas dataframe
df2 = df.set_index("api", drop = False)

#Set up the linear regression fit and prediction
class MyLinearRegression(LinearRegression):

    def fit(self, X, y):
        return super().fit(X, np.log(y))

    def predict(self, X):
        return np.exp(super().predict(X))

#Convert the text labels to binary data and fills NaNs with the average
mapper = DataFrameMapper([('H_or_V', LabelBinarizer()),
                          ('frac_proppant_type', LabelBinarizer()),
                          ('frac_slickwater', LabelBinarizer()),
                          ('frac_supplier', LabelBinarizer()),
                          ('operator_name', LabelBinarizer()),
                          ('operator_public', LabelBinarizer()),
                          ('primary_phase', LabelBinarizer()),
                          ('region_name', LabelBinarizer()),
                          ('reservoir_producing', LabelBinarizer()),
                          ('status', LabelBinarizer()),
                          ('parent_ticker', LabelBinarizer()),
                          ('well_name', LabelBinarizer()),
                          ('range', LabelBinarizer()),
                          ('township', LabelBinarizer()),
                         ],
                         default = Imputer(strategy='mean'))
```

*Figure 7*. First Half of the Code

The X variable represents all of the actual completion data from all of the columns (except the oil production column) for all fifty wells. The y variable represents all of the actual oil production data from all fifty wells. The Xtrain and ytrain data is actual completion and actual production data, respectively, that is used to train the model. This is done so the model will then in turn test the actual held-out values of the wells' actual completion and actual production, Xtest and ytest. The models are programmed to hold out about twenty percent of the data. This is customizable, if the data scientist wants to change the percentage. This is all setup in "train_test_split." The X data is then used to base off the predictions for the y_all values. Another way to put

it is that the line of code "y_all = model.predict(X)" runs the actual completion data, X, through Python in order to produce the predicted production values (y_all), based off the trained model. The trained model was created by the original actual X completion data and the original actual y production data. Then, this is once again performed, but now to the Xtest data, or the held-out completion data set, and that is used to predict what the new production values should be, y_holdout. It should be noted that y_holdout is different from ytest, as ytest is the actual held out production data, while y_holdout is the predicted production values based off the real held-out completion data, Xtest. This can be seen in Figure 8. The next paragraph summarizes the meaning of the graph produced below.

```
#Create the model
model = make_pipeline(StandardScaler(), PolynomialFeatures(1), MyLinearRegression(fit_intercept=True, normalize=True))

X = mapper.fit_transform(df2.drop('volume_oil_formation_bbls', axis=1));
y = df2['volume_oil_formation_bbls']

Xtrain, Xtest, ytrain, ytest = train_test_split(X, y, random_state=2)

model.fit(Xtrain, ytrain)

y_all = model.predict(X)
y_holdout = model.predict(Xtest)

#Create the graph
import matplotlib.pyplot as plt
max_plot = X.shape[1]
plt.plot(y, y_all, 'bo', ytest, y_holdout, 'ro')
plt.grid()
plt.legend(labels=["All", "Holdout"])
plt.xlabel('Actual Oil Production Data (bbls)')
plt.ylabel('Model Oil Production Data (bbls)')
plt.title('Testing the Fit of this Linear Regression Model', fontsize=14, fontweight='bold')

#Make a staight line with a slope of one
plt.plot([0,275000],[0,275000],'k')

plt.show()
```

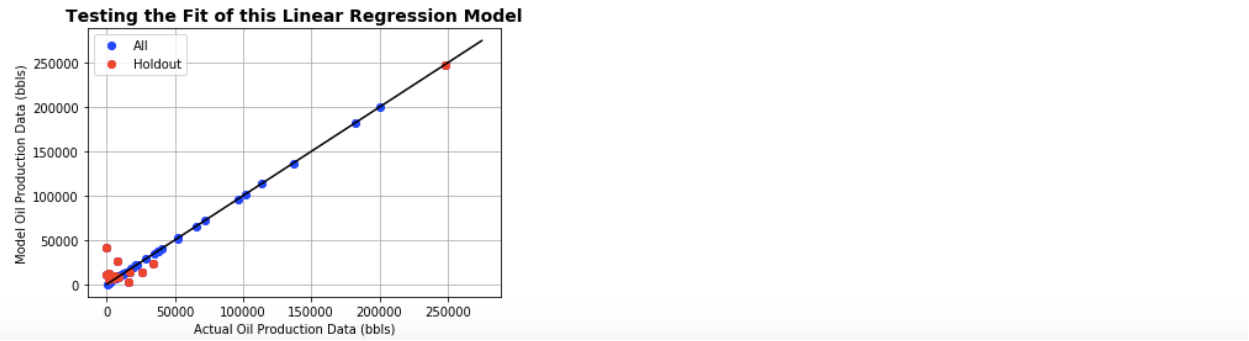


*Figure 8*. Second Half of the Code

After this, the code created a plot that combines both data sets of y_all (y-axis) versus y (x-axis) and y_holdout (y-axis) versus ytest (x-axis). In addition, a straight line with a slope of one was created to show how well the model predicted and fit the data. y_all versus y is the predicted production data based off all of the actual completion data versus the original actual production data. In theory, this should all on the straight line, as there is no new information, it is just testing the model's ability to fit the data. It is simply running all of the original data through Python based off of the model that was trained by the original data. y_holdout versus ytest is the new production values based off the held-out completion data versus the actual production values that were held out. The blue data points represent "All" of the data points, showing how well the model fit all of the actual production values, after running through the trained model. The red data points represent the "Holdout" data, showing how well the mode fit the new predicted production values based off the held-out data, after running through the trained model. In both cases, it is just different versions of the production values being graphed against each other. Ideally all of it would be on that straight line with a slope of one, as they would all ideally have a one to one, perfect ratio against the other. This is seen in Figure 9.

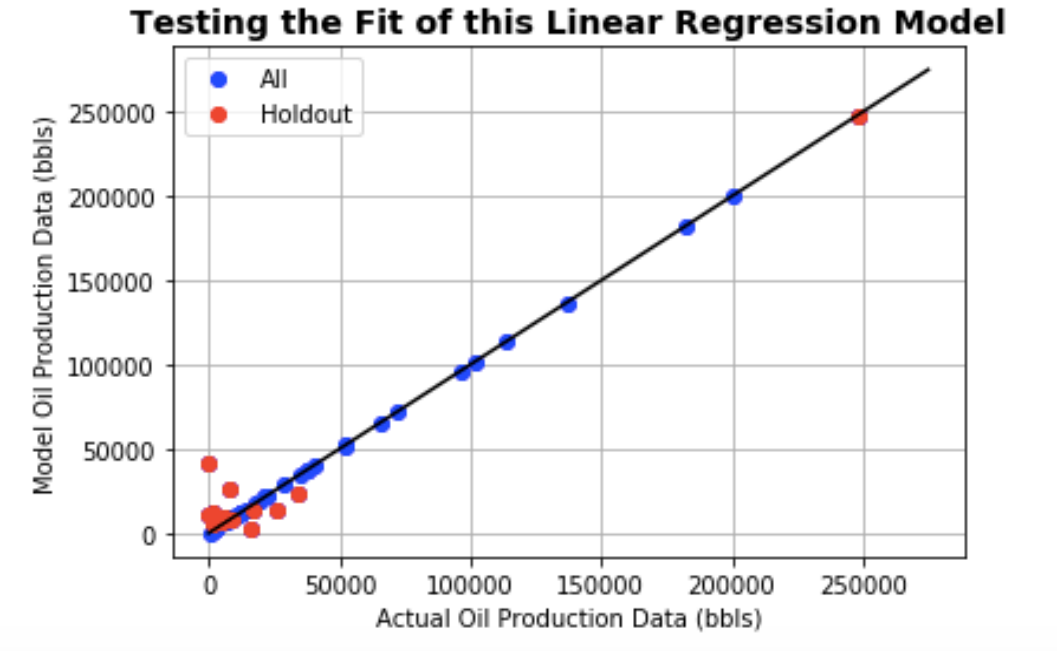


*Figure 9*. How Well the Model Predicts Data

The graph overall apparently shows all of the blue data points perfectly on the line and a lot of the red data points either on the line or very close to the line. This leads to the reasonable conclusion that Python trained the model very well and the predicted values are pretty close estimates to what real-life wells would produce. Therefore, the conclusion can be made that machine learning is a fast and an effective measure of predictions that can lend an extremely useful hand in predicting future oil wells' optimal barrels of production, given the completion and production data of previous wells in a specific area.

It should be noted that the number of fifty wells is not nearly enough data, not even three hundred is. Some amount – nearer thousands of wells – offers a much better data set. This plot would be more accurate if there were more wells included in this data set, as well as, if there was more information as how to handle the NaN values.

Most of the time, both of the above caveats are accounted for in real life machine learning situations in the oil and gas industry. However, the objective of this was to show just a very simple and generic version of the utility and potential behind this technology, greatly impacting large data sets, predicting new information, and exposing trends.

**Challenges & Considerations**

For decades now, the oil and gas industry has been gathering data and automating its systems. However, big data in the energy domain was only adopted recently, unlike other industries. Another caveat is that there simply is neither enough physical data nor velocity of data in this industry to be deemed necessarily as "big data." In addition, simple resistance to cultural change from industry leaders could hold this industry back. Another main concern is data security and the possibility of someone accessing private and sensitive data.

Dr. Gruss and Robert Skiebe agree that "the foremost challenge that most companies face is to gather actionable data" (2018). Quite a bit of data has only been gathered on an "'as-needed basis'" (Gruss, 2018). Due to this, only small amounts of information are available sometimes, which causes the collected data to be uncorrelated and non-actionable. This then results in a potential time-consuming and manual data collection analysis, which could prevent in solving time-sensitive issues (Gruss, 2018).

The main challenge is to gather data, bring it into a single platform where it can be analyzed and value generated, all in a short amount of time. Currently, if a company needs to solve an operational problem, they typically spend weeks or months to collect and aggregate the necessary data. The amount of time depends on the scope of the problem.

The other concerns are data security, data silos, and data discrimination. There is a great fear in companies accessing sensitive data. This data can range from learning too much about people's healthcare or income levels to learning about a business's future plans that would affect stock prices. Increasingly, people are asked to strike a balance between the amount of personal data that they divulge and the convenience that big data-powered apps and services offer (Marr, n.d. -a). Then once the data is divulged, the issue of entrusting the company, employees, or whoever, with that data becomes necessary.

The concept of data silos refers to data that comes from a separate database or set of data files that are not part of an organization's enterprise-wide data administration. Once a large volume of data is attained, it is difficult to clean it from the unnecessary information, and then merge it with the company's standard database, as it will come in a different format. Over the years, industries have come up with more solutions on how to solve cybersecurity problems and the methods to collect the right data and make it available to be analyzed in a timely manner. However, updating improvements in these solutions is a necessity to move towards enhanced future technology.

Once the data is collected and now many things are known, is it acceptable to discriminate against people based on this data? Banks already use credit scoring to decide who can borrow money and insurance also is heavily data-driven. As the use of big data increase, more data, people, companies, etc. will be analyzed and assessed in greater detail. Care must be taken such that this is not done in a way that contributes to making life more difficult for those who already have fewer resources and access to information (Marr, n.d. -a).

Facing up to these challenges is an important part of big data and must be addressed by the organizations that want to take advantage of big data. According to

Bernard Marr, "failure to do so can leave businesses vulnerable, not just in terms of their reputation, but also legally and financially" (n.d. -a).

Companies need to continuously allocate instructional skills and expertise throughout the organization in a cohesive, cost-effective manner. They need to discover unknown interactions and influences. According to Gruss and Skiebe, "traditionally, companies would hire subject matter experts (SMEs) across all their facilities around the globe and deploy them to maintain a full coverage of skilled support" (2018). However, this process is an expensive proposition.

Companies need to understand the end goal. Technology should always serve the business requirements. If company leaders integrate a solution without a well-thought-through strategy, they will leave little impacts on the productivity or growth of the organization, and potentially set progress back indefinitely. According to Gruss and Skiebe "simply gathering, analyzing, and visualizing data without empowering the enterprise to effectively change its operations on the basis of that data marginalizes the digitalization value proposition" (2018). There needs to be a willingness to adapt and welcome change within an enterprise's culture and processes. The business model that should lead and the technology should follow. Therefore, company leaders should employ a strategic approach, not an opportunistic approach. When implemented correctly, digitalization is such a powerful tool that can truly transform a business, causing it to be more aggressive amongst its competitors.

Another caveat to utilizing big data in the oil and gas industry is that there simply is neither enough physical data nor velocity of data in this industry to be deemed necessarily as "big data." Even though about a million well produce oil and / or gas in the United States today, that is nowhere near the number of people that use Facebook. About fifty-eight percent of adult Americans use it (Weise, 2015). There are

approximately 325 million in the United States, making about one hundred and ninety million adults on Facebook. This does not even include all of the people below the age of eighteen, who use Facebook. Compared to just one million wells, Facebook has one hundred and ninety times information than the oil and gas industry. However, it is important to remember that it is not only about how much data that a company, industry, person has, but what any of those do with the data is far more important. Yes, the oil and gas industry does not have nearly the data that Facebook and Google have access to, but that is not to say the energy industry's data is not important nor useful. Their data is still very applicable. This caveat is more so a warning to be mindful with the term "big data" in regard to the accessible volume of data in this industry. Big data nonetheless will significantly change the energy business in the future, as it already has in various other industries.

For decades now, the oil and gas industry has been gathering data and automating its systems. However, big data in the energy domain was only adopted recently, unlike other industries. According to ERM's digital innovation and advisory lead Peter Tickler, data science has not been applied to energy efficiency historically. Although it has been used in many industries, especially financial services, it has lagged in the energy efficiency sector (Big data, 2018). The main qualm holding the oil and gas industry back is its resistance to cultural change and the first companies that make the leap towards big data implementation will stand to gain a significant competitive edge.

## CONCLUSION

It is important to remember that the primary value from big data comes not from the data in its raw form but from the processing and analysis of it, and the insights, products, and services that emerge from the analysis. The sweeping changes in big data technologies and management approaches need to be accompanied by similarly dramatic shifts in how data supports decisions and product / service innovation (Big data, n.d.). Understanding the data can deliver tailored and sustainable outcomes that make quite a difference. With energy prices expected to remain elevated, managing energy costs with an end-to-end solution, underpinned by data analytics, can be the most cost-effective way to get bills down (Using energy, 2017).

It must also be kept in mind that if more coding is used, more opportunities for failure exist. If software engineers do not understand the problem they are trying to solve, and even worse, do not care to, then raw data fed through faulty software (which will only do what it is programmed to do) will lead to decisions being made on the basis of false or misleading models of reality. Renowned Dutch computer scientist Edsger Dijkstra wrote that the programmer "'has to be able to think in terms of conceptual hierarchies that are much deeper than a single mind ever needed to face before" (Somers, 2017). So even when people know how to code, the problem of what to code arises. Looking into the future, it is important that people will not be allowed to write programs, if they do not understand the concepts behind them (Somers, 2017). This creates a new kind of failure and opportunity, especially for the oil and gas industry. Value is generated in the blend of big data and specific subject matter expertise. There are not that many people that know how to combine computer science knowledge with information about the oil and gas industry. This now opens a whole

new sector of untapped business that can strengthen the industry in both the short and long terms.

Why is the courtship of the oil and gas industry and big data only starting now? Why are they not closer to "marriage"? Theories are as follows: first, there is a lack of training in this industry to conduct such tasks. Second, there is also a lack of data quality, as running the models is not necessarily the difficult part. The difficult part is getting the raw data primed for machine learning to work on it. There is not a lack of data. However, data is not yet viewed as an asset. Third, according to executives at Bazean, a company's constrained budgets limit big data's potential, as further acceleration will require more capital. Data managing and gathering still generally reside on IT budgets for most oil and gas companies, while the budget for data initiatives are generally around $500,000. Fourth, data rights also limit big data's potential, as only a select few can receive all of the necessary data and then only a select few of those few can create worthwhile data. Bazean's executives believe that cost of failure is high because a data scientist can accidentally blow things up and / or lose his / her job in a tough job market. Thus, there is little incentive to step outside the box. Overall, there needs to be a shift in culture that embraces big data and technology's evolution. There needs to be a commitment to dedicate time and interest to this within a company. Most likely this call to commitment will come from a top executive who drives the idea of big data, mainly because big data's costs and the problem of having to dynamically redefine targets and ways to hit them demand a level of cooperation that is well managed from the top down.

Big data nonetheless will significantly change the energy business in the future, as it already has in various other industries. The oil and gas industry has an opportunity to capitalize on big data analytics solutions and needs to encourage this.

Then the oil and gas industry must educate big data on the types of data the industry captures in order to utilize the existing data in faster, smarter ways that focus on helping find and produce more hydrocarbons, at lower costs in economically sound and environmentally friendly ways.

## REFERENCES


Big data & energy productivity drives the future of energy. (2018). *ERM Business Energy*. Retrieved from https://www.ermpower.com.au/post_powering_on/big_data/

Big data – what it is and why it matters. (n.d.). *SAS Institute Inc*. Retrieved from https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Farris, A. (2012). How big data is changing the oil & gas industry. *Analytics Magazine*. Retrieved from http://analytics-magazine.org/how-big-data-is-changing-the-oil-a-gas-industry/

Gruss A., & Skiebe, R. (2018). The evolution of energy analytics. *Energy CIO Insights*. Retrieved from https://energy-storage-system.energycioinsights.com/cxo-insights/the-evolution-of-energy-analytics-nwid-221.html

Jupyter. (2018). Project Jupyter. Retrieved from http://jupyter.org

Marr, B. (2018). The key definitions of artificial intelligence (AI) that explain its importance. *Forbes*. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/02/14/the-key-definitions-of-artificial-intelligence-ai-that-explain-its-importance/2/#7562597a5284

Marr, B. (n.d. -a). What is big data? A super simple explanation for everyone. *Bernard Marr & Co.* Retrieved from https://www.bernardmarr.com/default.asp?contentID=766

Marr, B. (n.d. -b). What is machine learning – a complete beginner's guide. *Bernard Marr & Co.* Retrieved from https://www.bernardmarr.com/default.asp?contentID=1140

Miklovic, D. (2017). Big data and predictive analytics in oil and gas. *General Electric*. Retrieved from https://www.ge.com/digital/blog/big-data-oil-and-gas

Somers, J. (2017). The coming software apocalypse. *The Atlantic*. Retrieved from https://www.theatlantic.com/technology/archive/2017/09/saving-the-world-from-code/540393/

Using energy data to deliver value. (2017). *ERM Business Energy*. Retrieved from https://www.ermpower.com.au/post_powering_on/using-energy-data-deliver-value/

VanderPlas, J. (2016a). Data manipulation with Pandas. In *Python Data Science Handbook* (Data manipulation with Pandas). Retrieved from https://jakevdp.github.io/PythonDataScienceHandbook/03.00-introduction-to-pandas.html

VanderPlas, J. (2016b). Introducing Pandas objects. In *Python Data Science Handbook* (Data manipulation with Pandas). Retrieved from https://jakevdp.github.io/PythonDataScienceHandbook/03.01-introducing-pandas-objects.html

Weise, E. (2015). Your mom and 58% of Americans are on Facebook. *USA Today*. Retrieved from https://www.usatoday.com/story/tech/2015/01/09/pew-survey-social-media-facebook-linkedin-twitter-instagram-pinterest/21461381/

## BIOGRAPHY

Michelle Tankimovich was born on August 28, 1995 in Los Angeles, California as a first generation American to Russian parents. In 2003, she moved to The Woodlands, Texas, a suburb outside of Houston. Michelle received a B.S. in Petroleum Engineering from the Hildebrand Department of Petroleum Engineering and a B.A. in Plan II Honors from the College of Liberal Arts. While in college, Michelle gained an interest in finance and pursued opportunities to combine it with the petroleum industry. She graduated in the spring of 2018 and will begin working with Bazean Corp. as an analyst and petroleum engineer. Outside of college, she plans to continue her love of traveling and constant learning.